High-dimensional Statistics

Pradeep Ravikumar
UT Austin

Outline

1. High Dimensional Data : Large p, small n
2. Sparsity
3. Group Sparsity
4. Low Rank

Curse of Dimensionality

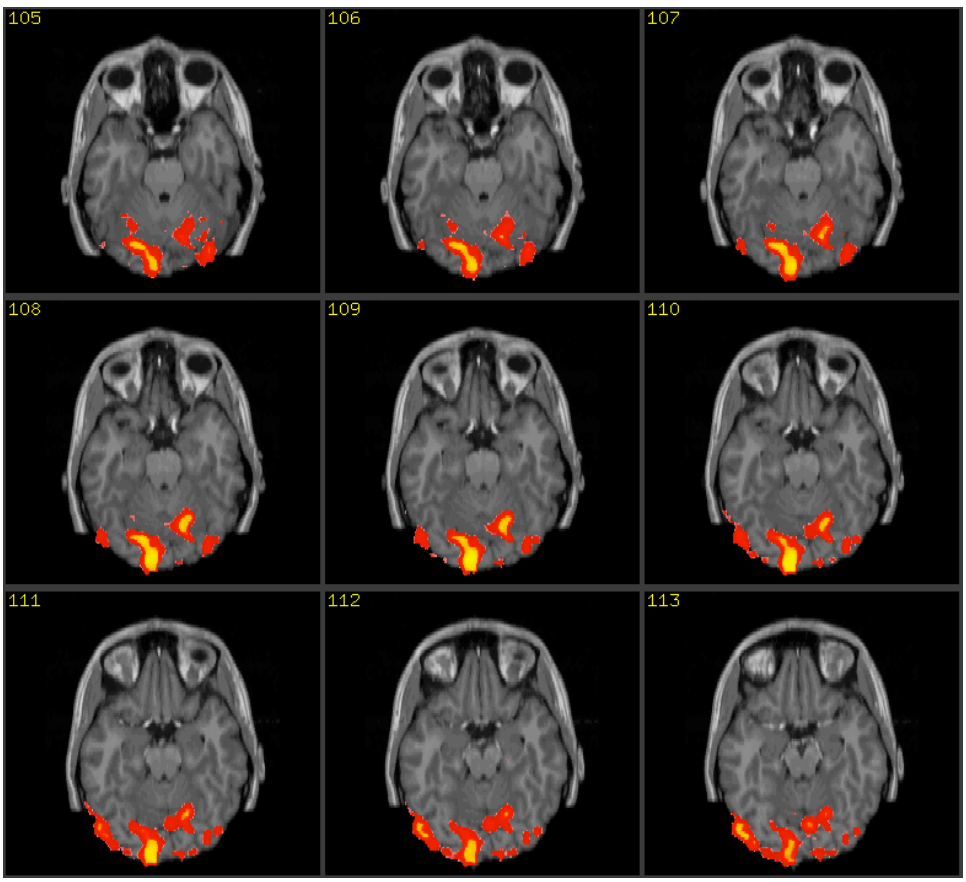
Statistical Learning: Given n observations from $p(X; \theta^*)$, where $\theta^* \in \mathbb{R}^p$, recover signal/parameter θ^*.

For reliable statistical learning, no. of observations n should scale exponentially with the dimension of data p.

What if we do not have these many observations?

What if the dimension of data p scales exponentially with the number of observations n instead?

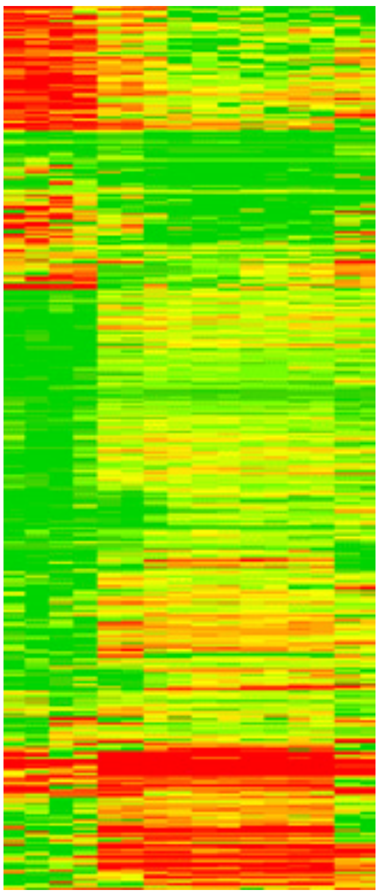
High-dim. Data: Imaging



Tens of thousands of "voxels" in each 3D image.

Don't want to spend hundreds of thousands of minutes inside machine in the name of curse of dim.!

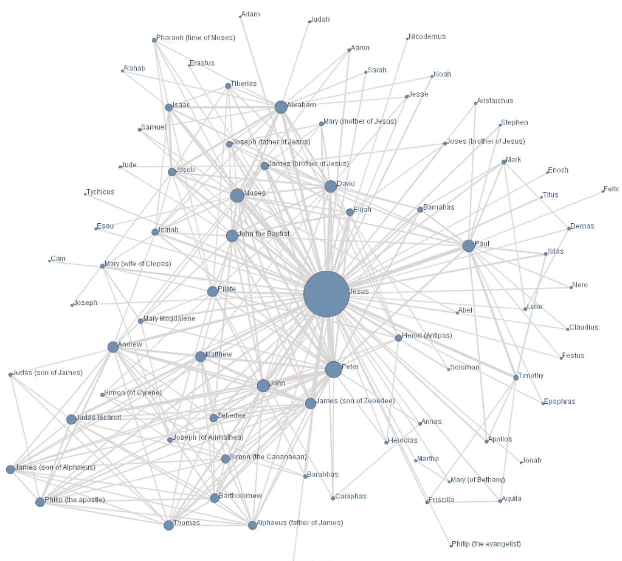
High-dim. Data: Gene (Microarray) Experiments



Tens of thousands of genes

Each experiment costs money (so no access to "exponentially" more observations)

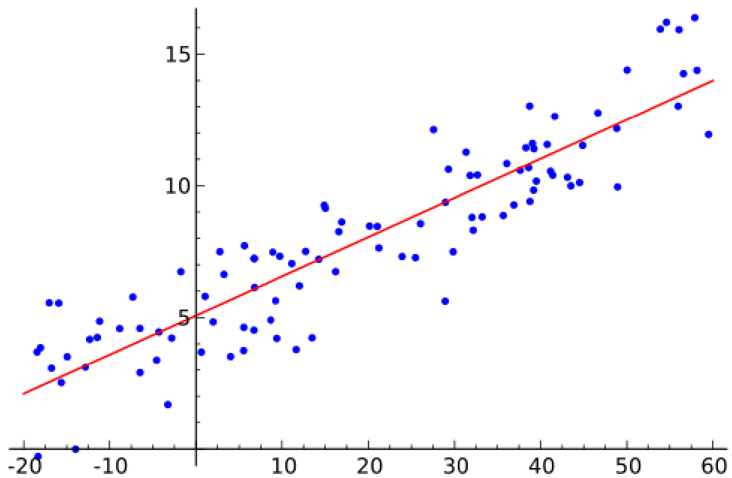
High-dim. Data: Social Networks



Millions/Billions of nodes/parameters

Fewer obervations

Linear Regression



Source: Wikipedia

$$Y_i = X_i^T \theta^* + \epsilon_i, \; i = 1, \ldots, n$$

Y : real-valued response

X : "covariates/features" in $\mathbb{R}^p$

Examples:

Finance: Modeling Investment risk, Spending, Demand, etc. (responses) given market conditions (features)

Epidemiology: Linking Tobacco Smoking (feature) to Mortality (response)

Linear Regression

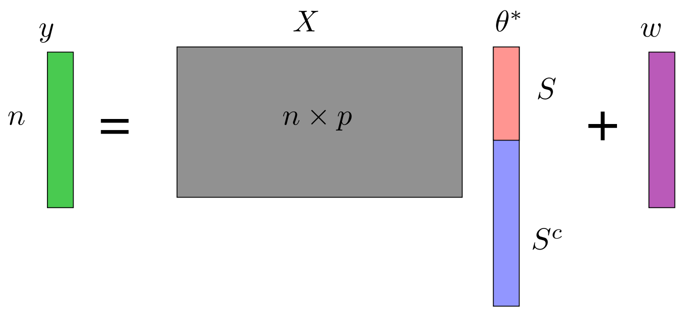
$$Y_i = X_i^T \theta^* + \epsilon_i,\ i = 1, \ldots, n$$

What if $p \gg n$?

Hope for consistent estimation even for such a high-dimensional model, if there is *some* low-dimensional structure!

Sparsity: Only a few entries are non-zero

Sparse Linear Regression



$\|\theta^*\|_0 = |\{j \in \{1, \ldots, p\} : \theta^*_j \neq 0\}|$ is small

Estimate a sparse linear model:

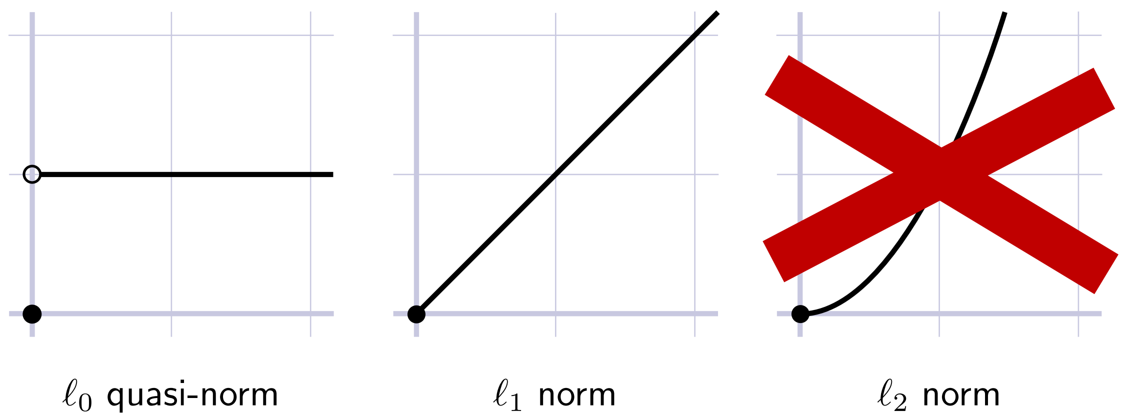
$$\min_{\theta} \|y - X\theta\|_2^2$$
$$\text{s.t. } \|\theta\|_0 \leq k.$$

ℓ_0 constrained linear regression!

NP-Hard : Davis (1994), Natarajan (1995)

Note: The estimation problem is non-convex

ℓ_1 Regularization



Source: Tropp 06

ℓ_1 norm is the closest "convex" norm to the ℓ_0 penalty.

ℓ_1 Regularization

Estimator: Lasso program

$$\widehat{\theta} \in \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \lambda_n \sum_{j=1}^{p}|\theta_j|$$

Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Meinshausen/Buhlmann, 2005; Candes/Tao, 2005; Donoho, 2005; Haupt & Nowak, 2006; Zhao/Yu, 2006; Wainwright, 2006; Zou, 2006; Koltchinskii, 2007; Meinshausen/Yu, 2007; Tsybakov et al., 2008

Equivalent:

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2$$
$$\text{s.t. } \|\theta\|_1 \le C.$$

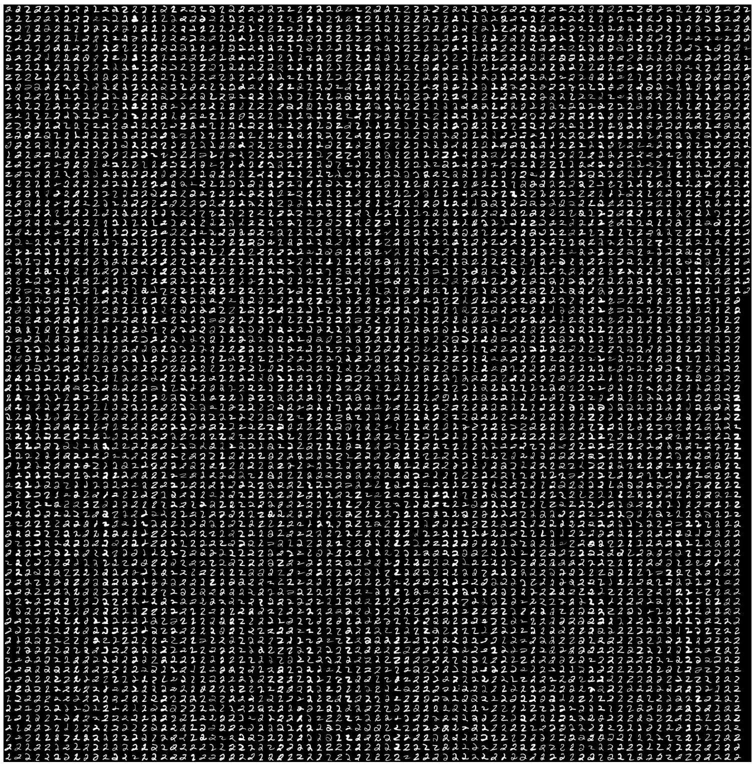
Group-Sparsity

Parameters in groups: $\theta = \big(\underbrace{\theta_1, \ldots, \theta_{|G_1|}}_{\theta_{G_1}}, \ldots, \underbrace{\theta_{p-|G_m|+1}, \ldots, \theta_p}_{\theta_{G_m}}\big)$

A **group** analog of sparsity: $\theta^* = \big(\underbrace{*, \ldots, *}_{\theta_{G_1}}, 0, \ldots, 0, \ldots\big)$

Only a few groups are active; rest are zero.

Handwriting Recognition



Data : Digit “Two” from multiple writers ; Task: Recognize Digit given a new image

Could model digit recognition for each writer separately, or mix all digits for training.

Alternative: Use group sparsity. Model digit recognition for each writer, but make the models share relevant features. (Each image is represented as a vector of features)

Group-sparse Multiple Linear Regression

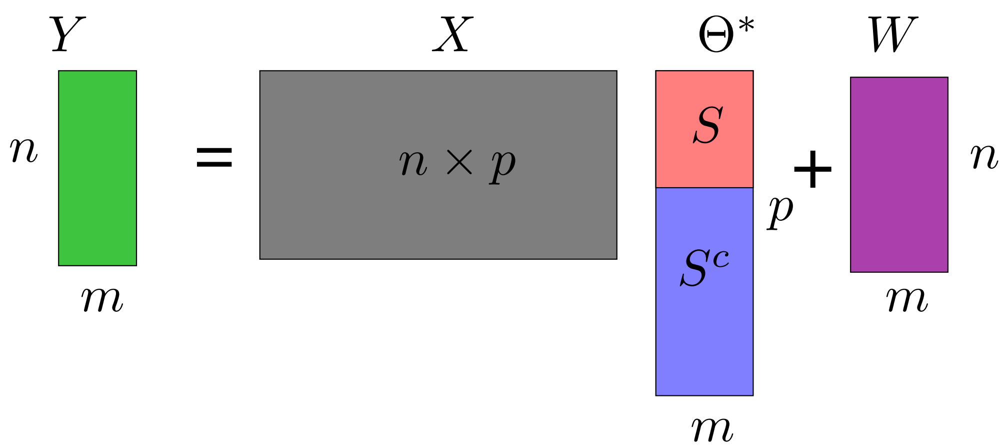
m Response Variables:

$Y_i^{(l)} = X_i^T \Theta^{(l)} + w_i^{(l)},\ i = 1, \ldots, n.$

Collate into matrices $Y \in \mathbb{R}^{n \times m}$, $X \in \mathbb{R}^{n \times p}$ and $\Theta \in \mathbb{R}^{m \times p}$:

Multiple Linear Regression: $Y = X\Theta + W$.

Group-sparse Multiple Linear Regression



Estimate a group-sparse model where rows (groups) of Θ^* are sparse:

$|\{j \in \{1, \ldots, p\} : \Theta^*_{j\cdot} \neq 0\}|$ is small.

Group Lasso

$$\min_{\Theta}\left\{\sum_{l=1}^{m}\sum_{i=1}^{n}(Y_i^{(l)} - X_i^T\Theta_{\cdot l})^2 + \lambda\sum_{j=1}^{p}\|\Theta_{j\cdot}\|_q\right\}.$$

Group analog of Lasso.

Lasso: $\|\theta\|_0 \to \sum_{j=1}^{p}|\theta_j|$.

Group Lasso: $\|(\|\Theta_{j\cdot}\|_q)\|_0 \to \sum_{j=1}^{p}\|\Theta_{j\cdot}\|_q$

(Obozinski et al; Negahban et al; Huang et al; ...)

Low Rank

Matrix-structured observations: $X \in \mathbb{R}^{k \times m}$, $Y \in \mathbb{R}$.

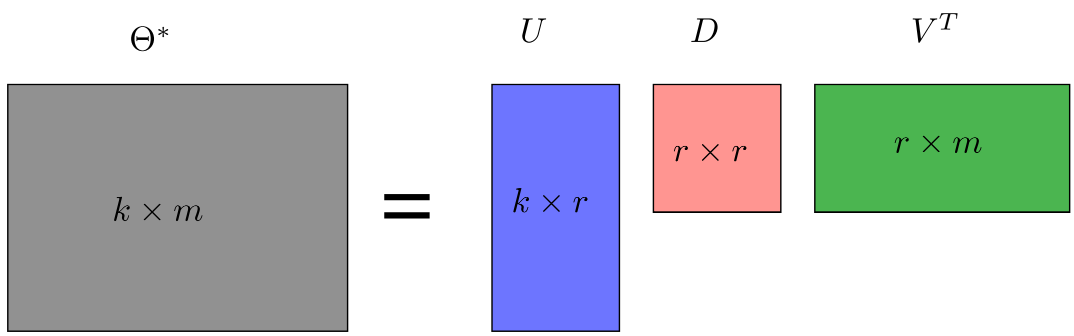
Parameters are matrices: $\Theta \in \mathbb{R}^{k \times m}$

Linear Model: $Y_i = \text{tr}(X_i \Theta) + W_i,\ i = 1, \ldots, n.$

Applications: Analysis of fMRI image data, EEG data decoding, neural response modeling, financial data.

Also arise in collaborative filtering: predicting user preferences for items (such as movies) based on their and other users' ratings of related items.

Low Rank



Set-up: Matrix $\Theta^* \in \mathbb{R}^{k \times m}$ with rank $r \ll \min\{k, m\}$.

Estimator:

$$\widehat{\Theta} \in \arg\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\langle X_i, \; \Theta \rangle\rangle)^2 + \lambda_n \sum_{j=1}^{\min\{k,m\}} \sigma_j(\Theta)$$

<u>Some past work:</u> Frieze et al., 1998; Achilioptas & McSherry, 2001; Srebro et al., 2004; Drineas et al., 2005; Rudelson & Vershynin, 2006; Recht et al., 2007; Bach, 2008; Meka et al., 2009; Candes & Tao, 2009; Keshavan et al., 2009

Nuclear Norm

Singular Values of $A \in \mathbb{R}^{k \times m}$: Square-roots of non-zero eigenvalues of $A^T A$.

Matrix Decomposition: $A = \sum_{i=1}^{r} \sigma_i u_i (v_i)^T$.

Rank of Matrix $A = |\{i \in \{1, \ldots, \min\{k, m\}\} : \sigma_i \neq 0\}|$.

Nuclear Norm is the low-rank analog of Lasso:

$\|A\|_* = \sum_{i=1}^{\min\{k,m\}} \sigma_i$.

High-dimensional Statistical Analysis

Typical Statistical Consistency Analysis: Holding model size (p) fixed, as number of samples goes to infinity, estimated parameter $\widehat{\theta}$ approaches the true parameter θ^*.

Meaningless in finite sample cases where $p \gg n$!

Need a new breed of modern statistical analysis: both model size p **and** sample size n go to infinity!

Typical Statistical Guarantees of Interest for an estimate $\widehat{\theta}$:

- Structure Recovery e.g. is sparsity pattern of $\widehat{\theta}$ same as of θ^*?

- Parameter Bounds: $\|\widehat{\theta} - \theta^*\|$ (e.g. ℓ_2 error bounds)

- Risk (Loss) Bounds: difference in expected loss